

Write up for “Relational similarity and the non-independence of features in similarity judgments” experimental simulations: statistics

Ruairidh McLennan Battleday
PhD COS
battleday@princeton.edu

1 Introduction

Five experiments are presented in Goldstone, Medin, and Gentner (1991) that examine the relationship between attributional and relational features in similarity judgements. We are interested in replicating the effect in Experiment 3, a similarity rating task that presents results averaged over subjects and six picture sets.

1.1 Materials and procedure

Participants were presented with pairs of stimuli (each composed of sets of objects), and made similarity comparisons between each stimulus ($\{A, B, C, D\}$) and the target ($\{T\}$), of the form “On a scale from 1 to 9, how similar are X and Y?”.

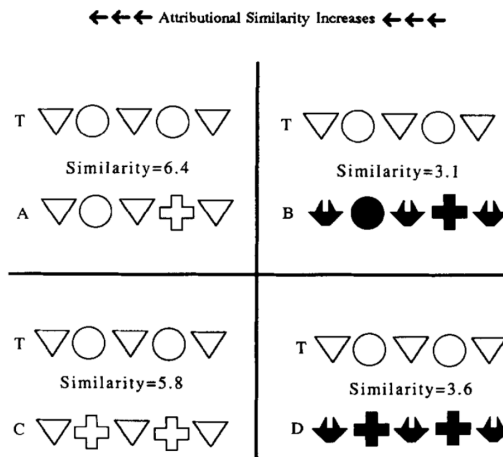


FIG. 4. Sample stimuli from Experiment 3. Feature independence is violated if subjects choose D as more similar to T than B , and choose A as more similar to T than C .

Figure 1: “On a scale from 1 to 9, how similar are $\{A, B, C, D\}$ and T ?”

Participants were assessed over six picture groups (set 2 shown in Figure 1). The objects varied in the features SHAPE and SHADE, and the relations used were SAME-SHAPE and SAME-SHADING. Attributional similarity was varied by changing the shape or shading of one or more of the objects such that no relation was disturbed.

1.2 Behavioural data

Average results for each picture set are presented in Figure 2. Weak support for MAX is found if $(D - B) - (C - A)$ is positive. The authors report that “the mean similarity ratings, collapsing over the six picture groups, show the same trends, yielding a value of .6 for $(D - B) - (C - A)$; this value is significantly greater than zero ($df = 5$, $t = 3.448$, $p < .02$)”. They also report that two of the individual stimulus sets have values significantly greater than zero (sets 2 and 3), but do not report the statistics; they also do not report whether these are significant under *post-hoc* correction for multiple testing, or whether they conducted any.

TABLE 1
Similarity Ratings with Respect to Target T in Experiment 3

| | Picture set | | | | | | Mean |
|---------------------|-------------|-------|-------|-------|-------|-------|------|
| | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 | |
| A | 5.8 | 6.4 | 5.5 | 5.9 | 4.3 | 6.7 | 5.8 |
| B | 3.4 | 3.1 | 4.4 | 4.0 | 2.1 | 6.0 | 3.8 |
| C | 5.2 | 5.8 | 4.9 | 6.2 | 4.4 | 6.8 | 5.6 |
| D | 3.6 | 3.6 | 4.5 | 4.8 | 2.1 | 6.8 | 4.2 |
| $(D - B) - (C - A)$ | .8 | 1.1 | .7 | .5 | -.1 | .7 | .6 |
| No. MAX Ss | 10 | 14 | 11 | 11 | 12 | 8 | 11 |
| No. MIN Ss | 5 | 3 | 5 | 6 | 9 | 3 | 5.2 |

Figure 2: Data from Exp. 3.

2 Pilot data

We conducted a pilot study on one of the six stimulus sets, in order to test whether the experiment would transfer to an online crowdsourcing-based paradigm and gain an estimate of the variability of subjects’ responses.

2.1 Stimuli

The stimulus set we used was based on that presented in Figure 1, with our replication shown in Figure 3.

2.2 Methods

We constructed the experiment using the javascript library jsPsych. English-speaking participants were recruited from Prolific (<https://app.prolific.co/>), an online crowdsourcing platform. They were paid \$0.8 for 5 minutes of participation, with no performance-related bonus.

The participants were consented, and viewed a set of instructions (Appendix A). They were then presented with all of the study stimuli twice, for three seconds each in random order, in order to familiarize them with the variability. After this, they were told that they would begin the rating task, which comprised clicking a button from 1-9 to rate the similarity of two stimuli (a target and a base) displayed side-by-side on the screen. After rating all of the stimulus pairs twice (once in each of left-right ordering) in random order, the participants were taken to a survey in which they recorded their age, gender, and years of education in mathematics.



Figure 3: Our replication of the stimuli from set 2 of experiment 3.

2.3 Results

The results from the pilot are recorded in Figure 4. The average time taken was 5 minutes and 5 seconds, with a range of 2 – 22 minutes. We recruited 33 participants, and 28 finished the study.

| | Ours | Original |
|--------------------|-------|----------|
| A | 5.43 | 6.4 |
| B | 1.61 | 3.1 |
| C | 4.59 | 5.8 |
| D | 2.16 | 3.6 |
| (D-B)-(C-A) | 1.39 | 1.1 |
| N_max | 1.00 | 14.0 |
| N_min | 0.00 | 3.0 |
| N | 28.00 | 29.0 |
| N_ind | 27.00 | 12.0 |

Figure 4: Results from the pilot experiment.

2.4 Statistics

The authors do not report the statistics or details of the test they conducted on this stimulus set, apart from stating that the result was significant. We can analyse whether $(D-B)-(C-A)$ significantly differed from zero using a t-test. Using the pilot mean of $\hat{\mu} = 1.39$ and standard deviation of $\hat{\sigma} = 1.701$, we obtain the following results: ($t = 4.324, p = 0.0002, \text{DOF} = N - 1 = 27$) (<http://powerandsamplesize.com/>).

3 Power analysis

We would like to extend the pilot study to consider six stimulus sets, and replicate the effect found in Goldstone et al. (1991) faithfully. The authors do not report enough information to guide a power analysis for the whole experiment, across all six stimulus sets, but we can try to derive the effective sample size needed to replicate these effects with a power analysis using only set 3 (Figure 2) as a guideline of the minimum significant effect size for individual stimulus sets. On the basis of the reported statistic, the authors used a one-sample two-tailed t-test to test whether the mean difference differed from zero. That is, they assume the test statistic under the null hypothesis follows a Student's t distribution with $N - 1$ degrees of freedom (formulae from Wikipedia):

$$t_{N-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{N}} \quad (1)$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{N - 1} \quad (2)$$

$$p(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu} \cdot \pi \cdot \Gamma(\frac{\nu}{2})} (1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}, \quad (3)$$

where $\nu = N - 1$ is the number of degrees of freedom.

Our goal is to calculate the sample size needed to replicate this effect given the significance level ($\alpha = 0.05$), as well as a power of 0.95 ($\beta = 0.05$). We know the mean for set 3 (0.7), as well as original sample size; however, we do not have access to individual-level data, and so cannot calculate the standard deviation.

3.0.1 Normal assumption

As Student's t distribution is known to converge to the Normal with increasing degrees of freedom, we could estimate the effective sample size based on a Normal assumption. Recall that the formula for the related Z-statistic is given as follows:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{N}} \sim \text{Normal}(0, 1). \quad (4)$$

We can arrive at an estimate for N by considering a two-sided power calculation for z-tests. Recall the following consideration of power (<https://cran.r-project.org/web/packages/distributions3/vignettes/one-sample-z-test.html>):

$$p(\text{reject } H_0 | \mu = \mu_A) = p(\bar{X} > \mu_0 + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} | \mu = \mu_A) + p(\bar{X} < \mu_0 + Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} | \mu = \mu_A), \quad (5)$$

where μ_A is our alternative hypothesis, at \bar{X} by construction. As $\bar{X} \sim \text{Normal}(\mu_A, \frac{\sigma^2}{N})$ by assumption, we can calculate these probabilities using simple rearrangements and the transformation given by the standard Normal distribution:

$$p(\bar{X} > \mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} | \mu = \mu_A) = p\left(\frac{\bar{X} - \mu_A}{\sigma/\sqrt{N}} > \frac{\mu_0 + z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} - \mu_A}{\sigma/\sqrt{N}}\right), \quad (6)$$

$$= p\left(Z > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{1-\frac{\alpha}{2}}\right). \quad (7)$$

Similarly,

$$p(\bar{X} < \mu_0 + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{N}} | \mu = \mu_A) = p\left(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\right). \quad (8)$$

Power at significance level α becomes as follows:

$$\text{Power} = p\left(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\right) + p\left(Z > \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{1-\frac{\alpha}{2}}\right). \quad (9)$$

Noting that only one of these terms contributes significantly if the null hypothesis has indeed been rejected (the parameter for alternative hypothesis can only more extreme in one direction), we can use algebra to recover N , as follows:

$$p\left(Z < \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}}\right) = 1 - \beta; \quad (10)$$

$$\therefore z_\beta = \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N}} + z_{\frac{\alpha}{2}} \quad (11)$$

$$\therefore N = \left(\frac{\sigma(z_\beta - z_{1-\alpha/2})}{\mu_0 - \mu_A}\right)^2. \quad (12)$$

Using $\alpha = 0.05$, $\beta = 0.05$, $\mu_0 = 0$, and $\mu_A = 0.7$, the effect found in the previous paper, we have the following approximation for N :

$$N = \left(\frac{\sigma \cdot (1.96 + 1.64)}{-0.7}\right)^2. \quad (13)$$

Although we do not know σ , we can use our pilot data to derive an empirical plug-in estimator ($\hat{\sigma} = 1.701$). With all of these estimates at hand, we find $N \approx 77$.

Finally, we can plot the values we obtained during the pilot, to see whether normality is a reasonable assumption. In this case, it does not seem so (but, the t distribution is relative robust anyway).

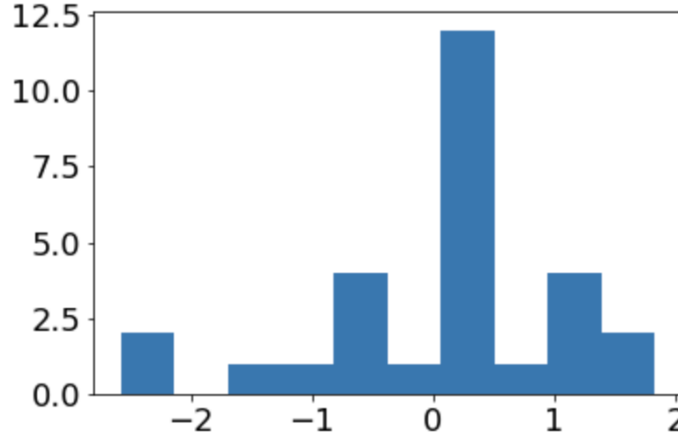


Figure 5: Data from pilot (D-B) - (C-A), z-scored.

3.0.2 t distribution

We could follow the same process using the t distribution directly. However, this runs into problems, as we shall see. Recall, the term S is:

$$S^2 = \frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N - 1} \quad (14)$$

This time, the relevant derivation is as follows (again WLOG restricting analysis to one tail):

$$p(\bar{X} < \mu_0 + t_{\alpha/2}^{(n-1)} \cdot S/\sqrt{N} | \mu = \mu_A) = p\left(\frac{\bar{X} - \mu_A}{S/\sqrt{N}} < \frac{\mu_0 + t_{\alpha/2} \cdot \frac{S}{\sqrt{N}} - \mu_A}{S/\sqrt{N}}\right), \quad (15)$$

$$= p\left(T < \frac{\mu_0 - \mu_A}{S/\sqrt{N}} + t_{\alpha/2}\right) \quad (16)$$

$$\therefore t_{\beta}^{(N-1)} = \frac{\mu_0 - \mu_A}{\sigma/\sqrt{N-1}} + t_{\alpha/2}^{(N-1)} \quad (17)$$

$$\therefore N = \left(\frac{\sigma(t_{\beta}^{(N-1)} + t_{1-\alpha/2}^{(N-1)})}{\mu_0 - \mu_A}\right)^2 + 1, \quad (18)$$

where we made use of the fact that $S/\sqrt{N} = \sigma/\sqrt{N-1}$.

The problem that immediately arises is that we need to know N to derive the degrees of freedom for the relevant t distribution, but we are trying to use that distribution to derive N in the first place. Some “authors” recommend beginning with the approximate N given by the Normal analysis above, and iterating until convergence (<https://stats.stackexchange.com/questions/146412/method-to-determine-the-sample-size-for-a-one-sample-t-test>).

Substituting $N = 77$, and the values above, we have:

$$“N” = \left(\frac{1.701 \cdot (1.67 + 2)}{-0.7}\right)^2 + 1, \quad (19)$$

$$= 80. \quad (20)$$

Obviously this is a discrepancy. However, we have convergence with one more iteration:

$$\text{DOF} = 79; N = 80. \quad (21)$$

We could also use the N from the last study to provide the initial degrees of freedom ($N = 29$). The convergence is as follows:

$$\text{DOF} = 28; N = 85; \quad (22)$$

$$\text{DOF} = 84; N = 80; \quad (23)$$

$$\text{DOF} = 79; N = 80. \quad (24)$$

3.0.3 Simulation

Another approach is to construct an empirical estimate of the distributions under different values of N using stochastic simulation, and test the significance and power of our effect size (e.g., following https://nickch-k.github.io/EconometricsSlides/Week_08/Power_Simulations.html).

That is, we sample $\{X_i\}_{i=1}^N \sim \text{Normal}(\mu_A, \hat{\sigma}^2)$, and derive the sufficient empirical statistics for our test (in this case, $(\tilde{\mu} = \bar{X}, \tilde{\sigma})$; $\tilde{S} = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{N-1}}$). We need to calculate the p-value and power of these data, and take the first N where the average of many simulations is in the range we want to find.

Recall, the formula for deriving the p-value of a two-sided one sample t-test is as follows:

$$t_{obs} = \frac{\tilde{\mu} - \mu_0}{\tilde{S}/\sqrt{N}}, \quad (25)$$

$$= \frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}/\sqrt{N-1}}, \quad (26)$$

$$\text{DOF} = N - 1, \quad (27)$$

and we look up $p_{val} \equiv p(|T| \geq |t_{obs}|) = 1 - p(T \leq t_{obs}) + p(T \leq -t_{obs})$ using a t-table / inverse cumulative density function (function).

And, the power can be found by the following:

$$t_{obs} = \frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}/\sqrt{N-1}}, \quad (28)$$

$$\text{DOF} = N - 1, \quad (29)$$

$$\text{Power} = p\left(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{S}/\sqrt{N}} + t_{\alpha/2}^{(\text{DOF})}\right) + p\left(T > \frac{\mu_0 - \tilde{\mu}}{\tilde{S}/\sqrt{N}} + t_{1-\alpha/2}^{(\text{DOF})}\right) \quad (30)$$

$$= p\left(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{\alpha/2}^{(\text{DOF})}\right) + p\left(T > \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{1-\alpha/2}^{(\text{DOF})}\right) \quad (31)$$

$$= p\left(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{\alpha/2}^{(\text{DOF})}\right) + \left[1 - p\left(T < \frac{\mu_0 - \tilde{\mu}}{\tilde{\sigma}/\sqrt{N-1}} + t_{1-\alpha/2}^{(\text{DOF})}\right)\right] \quad (32)$$

Running $M = 1000$ trials, and averaging across significance and power for each N , we obtain a Monte Carlo simulation estimate of $N \leftarrow \approx 115$ (see Figure 6).

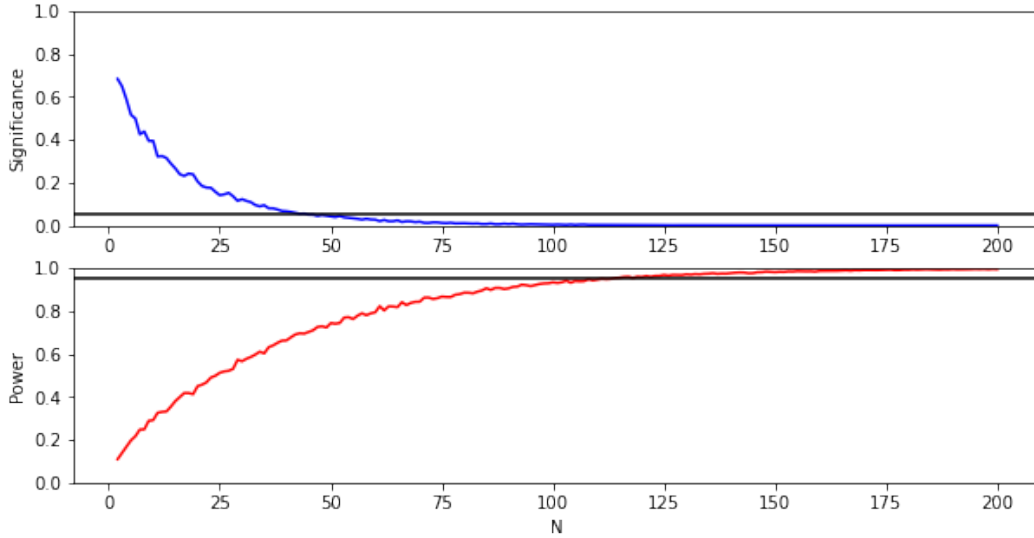


Figure 6: Simulation analysis of significance and power with N and the smallest effect size reported significant for a subset of stimuli in original study (set 3).

4 Replication experiment

We are now in a position to try to replicate the main effect found in Experiment 3 of Goldstone et al. (1991). This is a mean effect taken over all participants similarity judgements across six stimulus sets, with left-right order reversed.

4.1 Stimuli

The authors do not report the exact stimuli used in this experiment, apart from the stimulus set presented in Figure 1, above. We used their description of the other stimulus sets to select 5 others from their paper, which we reproduced using the `python` package `matplotlib` (Figures 7 and 8).

378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431

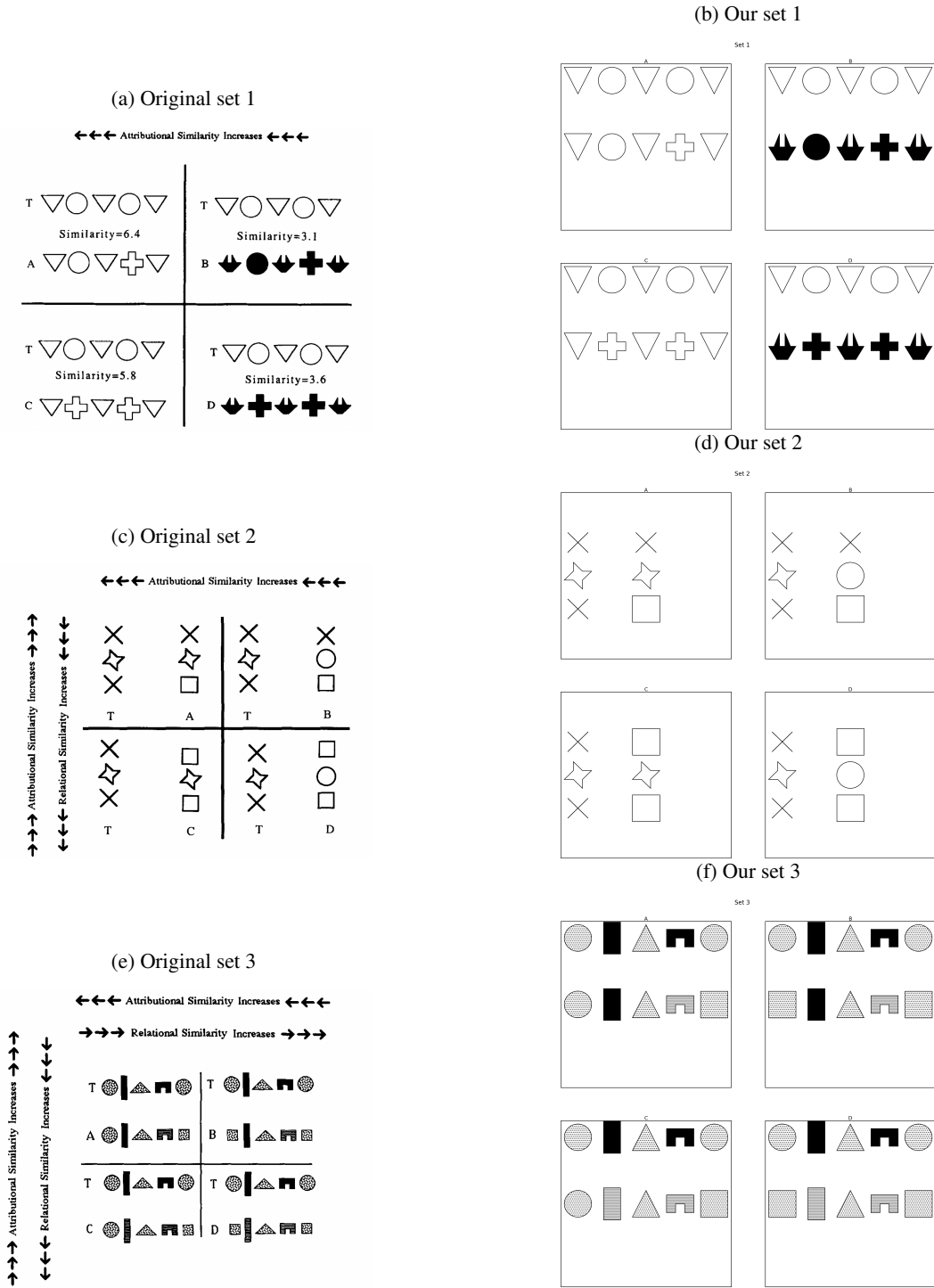


Figure 7: The first three stimulus sets.

432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485

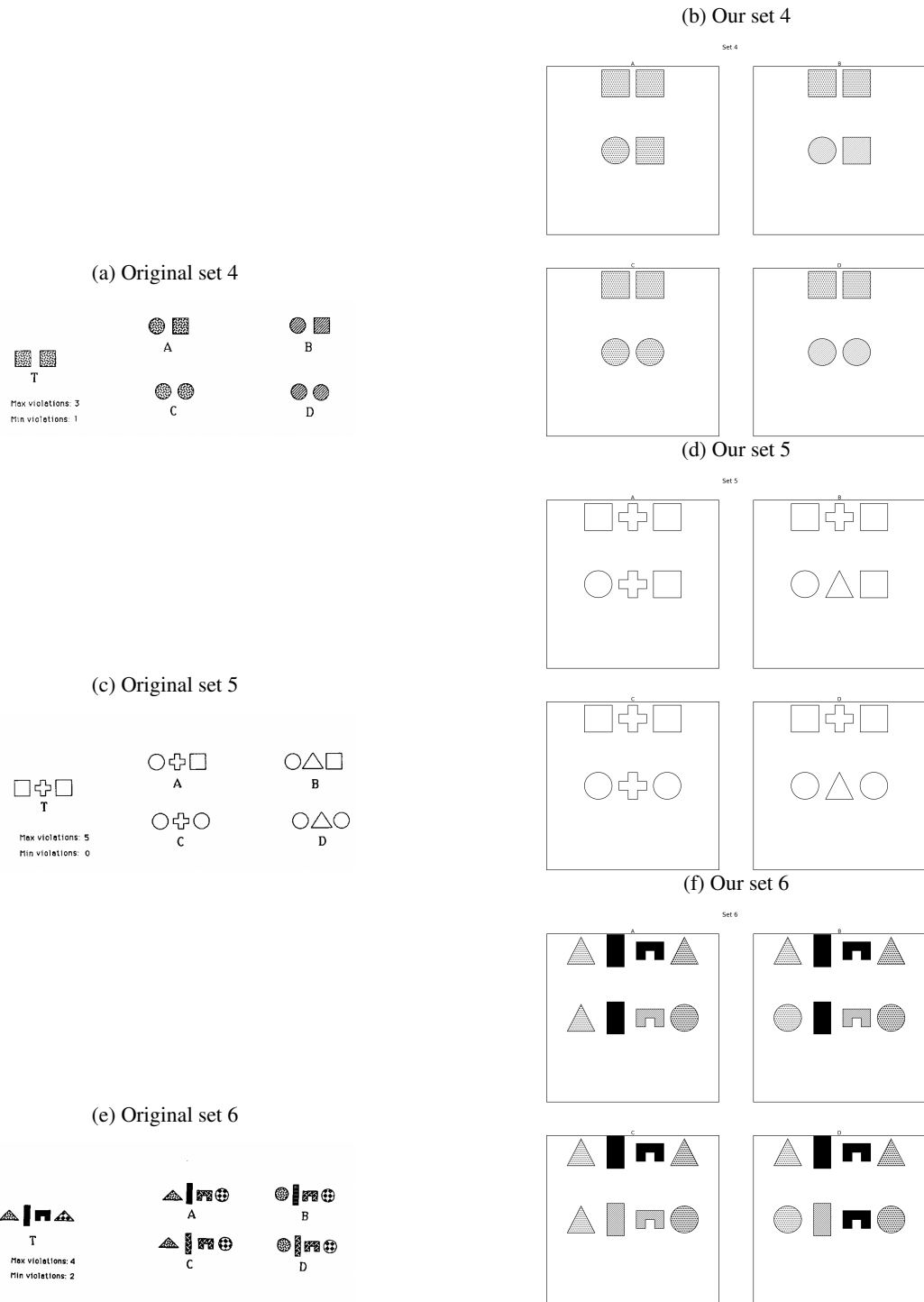


Figure 8: The final three stimulus sets.

4.2 Methods

The methods for the full experiment were designed to follow Goldstone et al. (1991) as closely as possible.

177 English-speaking participants were recruited from Prolific (<https://app.prolific.co/>), an online crowdsourcing platform. They were paid \$1.60 for 10 minutes of participation, with no performance-related bonus. The average completion time was 7 mins 14 seconds, with a range of 3 mins 30 seconds to 23 mins 20 seconds. 173 participants completed the study, but 5 of these did not pass our pre-registered exclusion criteria (trial time over 20 minutes or a single response chosen over 90% of the time). 166 remained, and we analyzed 160 of these at random as per our pre-registration.

Participants were consented, and presented with an introduction and set of instructions that described the trial stages and type of stimuli and judgements they would face (Appendix 2). During a familiarization phase, they were shown 25 of the stimuli for three seconds each, selected at random for each participant. Afterwards, they entered a ratings phase, in which a target and a base stimulus were presented side by side, and instructions appeared after one second asking them to rate the similarity of the stimuli from 1-9 by clicking an on-screen button (1 representing “not very similar”, 9 representing “highly similar”). The stimuli were presented in a random order, and each stimulus pair appeared twice in alternate left-right order (yielding 48 stimuli total). Afterwards, they were directed to an anonymous survey, where we collected age, gender, and years of math education.

4.3 Results

Results from the experiment are presented in Figure 9, below. The results are broken down across stimulus set, as per the original paper. Recall that a MAX result is given when the participant judges A as more similar to T than C is to T , and D more similar to T than B is to T . MIN results are the same, with the polarity reversed. We can also consider the aggregated results, shown in Figures 10 and 11.

4.4 Statistics

We can conduct a t test on the aggregate distribution as a first measure of whether the mean value of ‘(D-B)-(C-A)’ significantly differs from zero. The statistic will be based on our mean of $\hat{\mu} = 0.523$ and a standard deviation of $\hat{\sigma} = 0.865$. Recall, we are calculating the p-value based on the following formula:

$$t_{obs} = \frac{\tilde{\mu} - \mu_0}{\tilde{S}/\sqrt{N}}, \quad (33)$$

$$= \frac{\tilde{\mu} - \mu_0}{\tilde{\sigma}/\sqrt{N-1}}, \quad (34)$$

$$\text{DOF} = N - 1 \quad (35)$$

$$p_{val} \equiv p(|T| \geq |t_{obs}|) \quad (36)$$

$$= 1 - p(T \leq t_{obs}) + p(T \leq -t_{obs}). \quad (37)$$

From this calculation, we obtain ($t = 18.715$, $p = 3.669\text{e} - 67$, $\text{dof} = 959.0$).

We can also use a Bonferroni post-hoc correction to analyze each stimulus set, with a new significance level of 0.0083. Results are presented in Figure ?? . We can see the difference from zero in sets 1, 2, and 5 is highly significant. These sets are presented in Figure 13. Sets 3, 4, and 6, on the otherhand, are not significant. These sets are presented in Figure 14.

The authors also test whether an ordinal crossover between attributional and relational features occurred. That is, they test whether $\text{SIM}(A, T) > \text{SIM}(C, T)$, and $\text{SIM}(D, T) > \text{SIM}(B, T)$ at the aggregate level, separately. The appropriate test for each of these is a one-sided, two-sample, repeated measures (paired) t test. The t statistic in this case is (https://en.wikipedia.org/wiki/Student%27s_t-test#Independent_two-sample_t-test):

540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593

| | Set 1 | Set 2 | Set 3 | Set 4 | Set 5 | Set 6 | Mean |
|--------------------|--------|--------|--------|--------|--------|--------|--------|
| A | 5.62 | 4.37 | 5.32 | 4.82 | 5.16 | 4.78 | 5.01 |
| B | 1.92 | 2.74 | 4.76 | 3.25 | 2.91 | 4.21 | 3.30 |
| C | 5.08 | 3.34 | 4.94 | 4.71 | 4.76 | 4.41 | 4.54 |
| D | 2.13 | 2.52 | 4.51 | 3.43 | 3.30 | 4.18 | 3.35 |
| (D-B)-(C-A) | 0.76 | 0.81 | 0.14 | 0.29 | 0.79 | 0.35 | 0.52 |
| N_max | 1.00 | 5.00 | 3.00 | 2.00 | 7.00 | 3.00 | 3.50 |
| N_min | 0.00 | 1.00 | 1.00 | 3.00 | 0.00 | 2.00 | 1.17 |
| N | 160.00 | 160.00 | 160.00 | 160.00 | 160.00 | 160.00 | 160.00 |
| N_ind | 159.00 | 154.00 | 156.00 | 155.00 | 153.00 | 155.00 | 155.33 |
| Var_set | 2.84 | 3.75 | 3.21 | 2.58 | 3.95 | 4.03 | 3.39 |
| SD_set | 1.68 | 1.94 | 1.79 | 1.61 | 1.99 | 2.01 | 1.84 |
| SEM_set | 0.13 | 0.15 | 0.14 | 0.13 | 0.16 | 0.16 | 0.15 |

Figure 9: Results from the full experiment, broken down by stimulus set.

$$t = \frac{\bar{X}_D - \mu_0}{S_D / \sqrt{N}}, \quad (38)$$

where \bar{X}_D is the average of the difference between pairs, $\mu_0 = 0$, and $S_D = \sqrt{\frac{\sum_i (X_{Di} - \bar{X}_D)^2}{N-1}}$.

Recall, the p-value for a one-sided t-test is $p_{val} \equiv 1 - p(T \leq -t_{obs})$ if we expect t_{obs} to be greater than zero.

| | |
|--------------------|--------|
| A | 5.01 |
| B | 3.30 |
| C | 4.54 |
| D | 3.35 |
| (D-B)-(C-A) | 0.52 |
| N_max | 21.00 |
| N_min | 7.00 |
| N | 960.00 |
| N_ind | 932.00 |
| Var_all | 0.75 |
| SD_all | 0.87 |
| SEM_all | 0.03 |

Figure 10: Results from the full experiment, aggregated across stimulus sets.

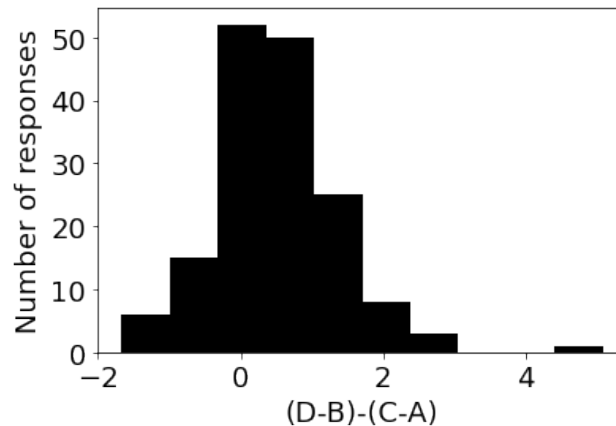


Figure 11: Results from the full experiment, aggregated across stimulus sets.

5 Appendix

5.1 Instructions for pilot experiment

"Welcome to the experiment. Press any key to begin."

"In this experiment, we will ask you to rate the similarity of different visual stimuli. In the training phase, we will present the stimuli one at a time. No response is required. In the testing phase, we will ask you to rate the similarity of two stimuli by clicking a button on the screen. Press any key to continue."

"In this training phase, we will show you the stimuli that we will be using in this experiment. No response is required. Press any key to continue."

| | t | p | DOF |
|--------------|---------|---------|-------|
| Set 1 | 5.66108 | 0.00000 | 159.0 |
| Set 2 | 5.28896 | 0.00000 | 159.0 |
| Set 3 | 0.96736 | 0.33483 | 159.0 |
| Set 4 | 2.28030 | 0.02392 | 159.0 |
| Set 5 | 5.03720 | 0.00000 | 159.0 |
| Set 6 | 2.17847 | 0.03084 | 159.0 |

Figure 12: Results from the full experiment, aggregated across stimulus sets.

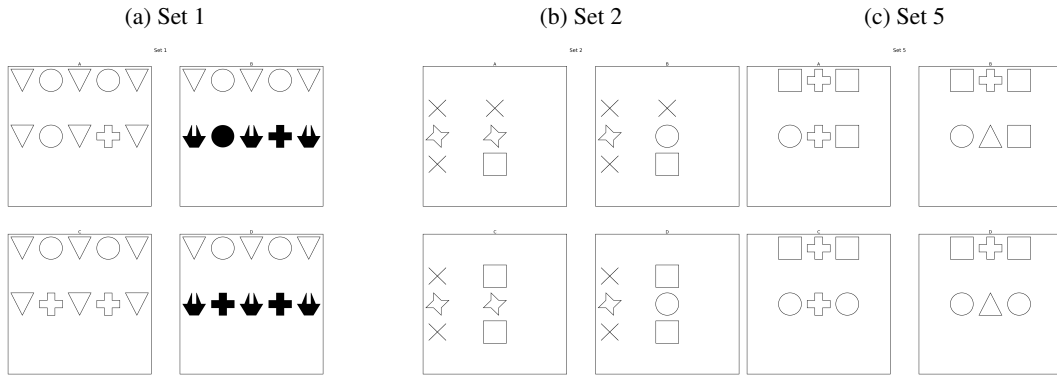


Figure 13: The significant stimulus sets.

"In this testing phase, two stimuli will appear in the center of the screen. We will ask you to press a button from 1 (not very similar) to 9 (highly similar) to indicate how similar the stimuli are. Press any key to continue."

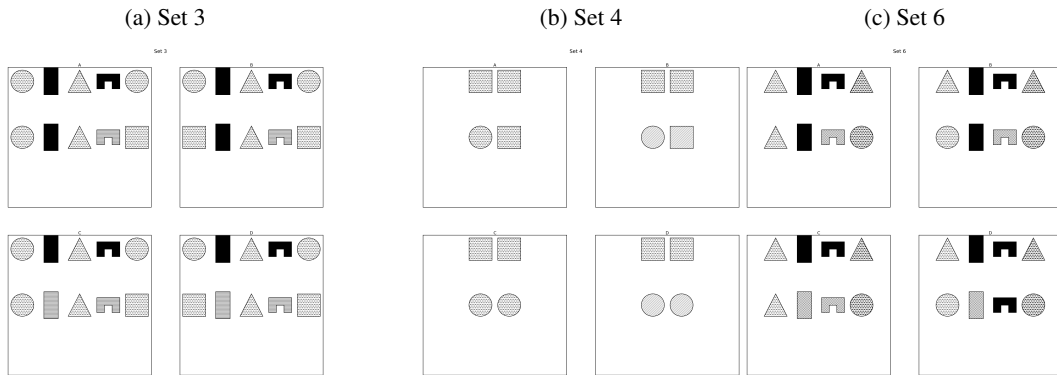


Figure 14: The nonsignificant stimulus sets.

"In this experiment, two stimuli will appear in the center of the screen. Press a key from 1 (not similar) to 9 (very similar) to indicate how similar the stimuli are. Press any key to continue."

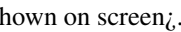
"Thank you for finishing! We have automatically recorded your Participant ID. Press any key to advance to an anonymous survey, which we are using for piloting."

"Thank you for finishing! Your answers will help us develop these experiments. Press any key to finish."

"Press any key to complete the experiment. Thank you."

5.2 Instructions for full experiment

"Welcome to the experiment. Press any key to begin. There may be a short delay as the experiment loads."

"In this experiment, we are interested in understanding how people make similarity judgements. We will ask you to rate the similarity of different pictures. Each picture is a collection of 2-5 shapes of different shading, and will look something like this: .

In the familiarization phase, we will present the pictures one at a time so that you can become familiar with them. Simply observe the picture, and when the instruction appears press any key to move on to the next one.

In the rating phase, we will present the pictures two at a time, and ask you to rate the similarity of two pictures by clicking a button on the screen. Press any key to continue."

"In this familiarization phase, we will show you the pictures that we will be using in this experiment one at a time so that you can become familiar with them. There will be 25 pictures. Simply observe the picture, and when the instruction appears press any key to move on to the next one. Press any key to continue."

"In this rating phase, two pictures will appear in the center of the screen at the same time.

We will ask you to rate the similarity of the two pictures. When the instruction appears, click a button to record your rating. Click the "1" key if the pictures are not very similar Click the "9" key if the pictures are highly similar. Click the other number keys if the similarity is in between these.

Press any key to continue."

"How similar are these pictures? Click a button from 1 (not very similar) to 9 (highly similar)."

"You have finished! Thank you. We have automatically recorded your Participant ID. Press any key to advance to an anonymous survey, which we are using to collect demographic data."

"This is a short survey to help us design further experiments. We would be most grateful if you could answer to the best of your ability. There is an option to not provide an answer to each question. "

"Thank you for finishing the survey! Your answers will help us develop these experiments.

Press any key to finish."

References

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive psychology*, 23(2), 222–262.